FEBRUARY 28, 2022

INDEPENDENT STUDY OF DIFFERENTIAL ITEM ANALYSIS IN EDMENTUM'S ITEM BANK

EDMETRIC LLC

EXECUTIVE SUMMARY

The purpose of this study is to investigate the presence of differential item functioning (DIF) in the Edmentum item bank. Differential item functioning statistics are used to quantify differences in item performance between two groups after controlling for examinees' overall achievement level. Typically, these statistics are used to examine performance differences between relevant subgroups of equal ability, such as males and females, on an assessment. In this study, the presence of DIF was investigated using the Mantel–Haenszel (MH) procedure.

The MH procedure has a straightforward implementation and enables the use of the classification system established by Educational Testing Service (Zwick & Ercikan, 1989). This classification system has been used widely in K-12 assessment, and it separates items into differing levels of DIF including negligible DIF (A-level), moderate DIF (B-level), and large DIF (C-level). Items flagged with B- and C-level DIF are the items of concern because they indicate that students in the groups of interest perform differentially on the item. This study examined the impact of four grouping variables, including gender, race, socioeconomic status (SES), and pandemic effect.

Despite the large number of items in the Edmentum's item bank, no items were flagged for B- or C-level DIF. DIF procedures are one means of examining test construct. When tests are beset with items flagged for DIF, this indicates that the construct may be measured differently between groups. The lack of DIF items in the current study provides evidence that the items in Edmentum's item bank are measuring student achievement from different groups in a similar manner. It provides some evidence that Edmentum's items are fair for different groups.

TABLE OF CONTENTS

TABLES

OVERVIEW

The purpose of this study is to investigate the presence of differential item functioning (DIF) in the Edmentum item bank. Differential item functioning statistics are used to quantify differences in item performance between two groups after controlling for examinees' overall achievement level. Typically, these statistics are used to examine performance differences between relevant subgroups of equal ability, such as males and females, on an assessment. DIF occurs when the probability of success on an item systematically differs for students from different groups with the same underlying true ability. Conversely, an item is considered free of DIF when different groups of students with similar ability levels have equal probability of success on that item. Moreover, uniform DIF is present when the DIF is in the same direction across the ability level continuum, whereas non-uniform DIF occurs when the direction of DIF is different depending on the ability level. In other words, the item consistently favors one group across all ability levels in uniform DIF; however, the direction of DIF changes at different locations of the ability levels in non-uniform DIF. In this study, the presence of DIF was investigated using the Mantel–Haenszel (MH) procedure which is one of the most (if not the most) widely-used procedure to evaluate DIF (Clauser & Mazor, 1998). Four different grouping variables were investigated in this study, including gender, race, poverty, and students affected by the pandemic.

MANTEL HAENSZEL

The MH chi square statistic was introduced in 1959 and popularized to detect the DIF in the late 1980s and early 1990s (Holland & Thayer, 1988; Dorans & Schmitt, 1993). The appeal of the MH procedure likely stems from its ease of implementation and the classification system established by Educational Testing Service (ETS; Zwick & Ercikan, 1989). This classification system allows researchers to easily interpret the severity of DIF by its effect size and to have content specialists further investigate those items with moderate and large DIF.

The MH procedure has been widely studied, but a detailed literature review is beyond the scope of the current study. Here, some relevant themes in the DIF literature are highlighted. First, the MH procedure has been widely used as a comparative method in DIF studies of Lord's chi square (Huang, Church, & Katigback, 1997); SIBTEST (Fidalgo, Ferreres, & Muñiz, 2004; Finch, 2005; Gierl, Jodoin, & Ackerman, 2000; Narayanon & Swaminathan, 1996; Roussos & Stout, 1996); logistic regression (Rogers & Swaminathan, 1993; Narayanon & Swaminathan, 1996; Gierl, Jodoin, & Ackerman, 2000; Güler & Penfield, 2009; Swaminathan & Rogers, 1990; Wiberg, 2009); and MIMIC models (Finch, 2005). Results from these studies are mixed. In general, the MH procedure performed comparably to these methods, except in detecting items with non-uniform DIF.

Second, researchers have tried to find ways to improve the MH procedure's ability to detect non-uniform DIF. Rogers (1989) as cited in Mazor, Clauser, and Hambleton (1994) showed that MH was able to detect some non-uniform DIF, but only for easy or difficult items and not items in the middle of the ability range. This is not surprising because of the way the MH procedure is calculated. Negative differences in ability in one part of the theta scale may be masked by positive differences in ability in another part of the scale (Mazor, Clauser, & Hambleton, 1994). Modifications to the MH method have been suggested to improve its ability to detect non-uniform DIF. One set of researchers proposed an unsigned MH only to conclude that the unmodified MH performed equally well (Nohoon, Davison, & Davenport, 1997). Several

researchers have recommended estimating the MH statistic separately for the upper and lower half of the ability range to improve its ability to detect non-uniform DIF (Mazor, Clauser, & Hambleton, 1994; Fidalgo, Mellenberg, & Muñiz, 2000; Maranón, Garcia, & Costas, 1997). Fidalgo, Mellenberg, and Muñiz (2000) caution that this method is not robust for large sample sizes.

Third, researchers study how well MH correctly detects DIF when manipulating ability distribution, sample size, and test length. Most of the studies reported that MH performed better (lower Type I error rates or inflation of DIF estimates) under conditions where the mean ability of the focal and reference groups was equal (Clauser, Mazor, & Hambleton, 1993; Fidalgo, Ferreres, & Muñiz, 2004; Roussos & Stout, 1996; Zwick, 1990); larger sample sizes were used (Gierl, Jodoin, & Ackerman, 2000; Narayan & Swaminathan, 1994); and longer tests (greater than 20 items) were used (Monahan & Ankenmann, 2005; Fidalgo, Ferreres, & Muñiz, 2004).

METHOD

This study involved completing a series of DIF analyses using student responses to items from Edmentum's item bank and delivered through Edmentum's Exact Path Diagnostic Computerized Adaptive Test. All the analyses were performed using difR (Magis, Beland, Tuerlinckx, & De Boeck, 2010) and tidyverse (v1.3.0; Wickham et al., 2019) packages in R. First, the data cleaning procedures are described. Next, the design for the DIF analyses using student responses from nine item-level datasets (i.e., three different subjects and three different academic years) is presented. Specifically, the presence of DIF was investigated using the Mantel–Haenszel (MH) procedure (Clauser & Mazor, 1998). This method allowed for detecting uniform DIF without requiring an item response theory model. The MH procedure has a straightforward implementation and enabled the use of the classification system established by Educational Testing Service (Zwick & Ercikan, 1989). The ETS DIF classification system is based on the MH delta statistic (*MH Delta*). This statistic,

$$MH\ Delta = -2.35(\ln(\alpha)),$$

where α is the MH odds ratio estimate (Mantel-Haenszel, 1959). This classification system allowed researchers to apply widely accepted rules on the severity of DIF by its effect size (small, moderate, or large). Specifically,

- Moderate DIF (B-level): Significant MH chi-square statistic (p<0.05) and $1.0 \leq |MH\ Delta| < 1.5$
- Large DIF (C-level): Significant MH chi-square statistic (p<0.05) and $|MH\ Delta| \geq 1.5$
- Negligible DIF (A-level): Otherwise.

Items flagged for C-level DIF should not be used on operational assessments. Items flagged for B-level DIF may be used if there are no other items available and the B-level items have passed a bias review. A-level DIF items may remain in the general pool. The difference between the item-level performance of the different groups is negligible, and the items may be treated as if there are no detectable differences.

In statistical testing, power represents the probability of correctly rejecting the null hypothesis. Statistical significance and statistical power depend on the effect size and sample size. Moreover, the results of a statistical test using a large sample size may not be statistically significant, while it may be significant in

terms of the effect size. Therefore, it is advisable to examine the results in the context of both statistical significance and effect size (Cohen, 1988). In DIF analysis, the interpretation of the results using both statistical significance and effect size will help reduce Type I errors.

CURRENT STUDY DATA

For the current study, Edmentum provided Exact Path data files for three different subjects (i.e., language arts, mathematics, and reading) from three different school years (i.e., 2018-2019, 2019-2020, and 2020-2021). Specifically, these datasets included students' responses to specific items in three subjects with three school years (i.e., 3x3 = 9 datasets). In addition, two demographic files (i.e., student-level and account-level) were also provided. These datasets were used to obtain grouping variables used in the DIF analyses. Moreover, the data files included three test-level diagnostic data files (one per school year), one item metadata file, and one item response frequency table based on the 2020-2021 school year.

DATA CLEANING

According to documentation provided to EdMetric by Edmentum, item-level and test-level datasets had already been cleaned according to some of Edmentum's standard cleaning processes (e.g., test event is complete; test event was completed within 14 consecutive days with no more than three unique days of actively responding to items; test event did not receive a flag indicating the student may have been rushing). In addition, EdMetric used the following rules to clean the datasets:

RULE 1. Nine item-level datasets were converted to student by item data frames separately. If a student has more than one response to a certain item, then the response with the earliest date was used in the analyses.

RULE 2. Field test items were excluded from the analyses.

RULE 3. The retired items were excluded.

RULE 4. After the grouping variables were identified, the distribution of the responses in each item with respect to the group variable were examined. If an item had fewer than 50 responses in each category of the grouping variable (i.e., 50+50 = 100 in total), that item was excluded from the analyses. Typical DIF analyses require at least 100 per group; however, a lower threshold was used to retain a greater number of items. It should be noted that using a small sample size in DIF analysis may lead to a greater Type I error; however, this is more often observed in more complicated models, such as 3PL IRT model.

After implementing the rules, the original datasets were converted to student-by-item data frames. Then, these converted response datasets were merged with student-level and account-level demographic datasets.

Table 1 lists the data sources along with the number of rows. It should be noted that the number of rows does not represent the number of unique students in these datasets. Each row represents a response to an item. The number of the responses to the items varies item by item because of the characteristic of the test (i.e., the nature of computerized adaptive testing).

TABLE 1. ITEM-LEVEL STUDENT RESPONSE DATA BY SUBJECT AND SCHOOL YEAR (AFTER APPLYING RULES 1 – 3)

	Before Cleaning		After Cleaning	
Dataset	Number of Rows	Number of Items	Number of Rows	Number of Items
2018-2019 Math	20,991,673	4,227	17,304,653	3,775
2019-2020 Math	16,469,068	3,360	14,919,538	3,121
2020-2021 Math	41,982,727	3,560	38,500,996	3,380
2018-2019 Language Arts	12,165,570	2,291	10,137,033	1,970
2019-2020 Language Arts	10,620,173	2,541	9,439,888	2,302
2020-2021 Language Arts	28,779,453	2,650	26,090,544	2,395
2018-2019 Reading	17,593,558	2,468	14,205,736	2,199
2019-2020 Reading	13,566,960	2,557	11,761,014	2,383
2020-2021 Reading	38,227,514	2,498	33,373,122	2,394

GROUPING VARIABLES

This study examined the impact of four grouping variables, including gender, race, socioeconomic status (SES), and the pandemic effect. The gender variable was obtained from student-level demographic data file. The race and SES variables were obtained from account-level demographic data file.

Gender. The gender variable was available in the student-level data set for almost half of the students (there is no gender indicated for the remaining students). Table 2 shows that the data were fairly evenly split between males and females, with slightly under half assigned female and slightly over half identified as male.

TABLE 2. PERCENTAGE OF STUDENTS BY GENDER, YEAR, AND SUBJECT AREA (AFTER DATA CLEANING)

Dataset	N Count	Female	Male
Overall (%49 is missing)	16,155,811	25%	26%
2018-2019 Math	127,991	48%	52%
2019-2020 Math	101,038	48%	52%
2020-2021 Math	315,883	49%	52%
2018-2019 Language Arts	69,955	47%	53%
2019-2020 Language Arts	63,562	48%	52%
2020-2021 Language Arts	234,393	49%	51%
2018-2019 Reading	105,238	48%	52%
2019-2020 Reading	88,996	48%	52%
2020-2021 Reading	310,860	49%	51%

Table 3 shows the number of items with fewer than 100 or 200 students in the focal group (males) and the reference group (females). As shown in Table 3, very few items had fewer than 100 students in the focal or reference group.

TABLE 3. THE NUMBER OF ITEMS WITH FEWER THAN 100 AND 200 IN FOCAL AND REFERENCE GROUPS FOR GENDER (AFTER DATA CLEANING)

Dataset	Number of Items	GENDER			
		Focal < 100	Focal < 200	Reference < 100	Reference < 200
2018-2019 Math	3,775	7	228	12	276
2019-2020 Math	3,121	0	149	6	165
2020-2021 Math	3,380	0	11	3	15
2018-2019 Language Arts	1,970	8	83	29	85
2019-2020 Language Arts	2,302	0	99	12	126
2020-2021 Language Arts	2,395	0	0	0	0
2018-2019 Reading	2,199	3	56	3	77
2019-2020 Reading	2,383	2	79	11	107
2020-2021 Reading	2,394	0	1	0	1

Note. Male is focal group.

Table 4 shows the distribution of sample sizes (focal and reference combined) for the DIF analyses conducted using the gender variable. The minimum sample size for 2018-2019 mathematics analyses was 157 students in the combined reference and focal categories. The median sample sizes were highest for the 2020-21 school year and lowest for the 2019-2020 school year.

TABLE 4. DESCRIPTIVE SAMPLE SIZES IN DIF ANALYSIS USING GENDER (AFTER DATA CLEANING)

Dataset	Minimum	1st Quarter	Median	Mean	3rd Quarter	Max	# Total Items
2018-2019 Math	157	890	2,780	3,422	5,114	18,702	2,932
2019-2020 Math	168	922	1,893	2,394	3,159	14,899	3,040
2020-2021 Math	186	2,676	5,500	6,747	8,964	43,547	3,377
2018-2019 Language Arts	155	1,441	2,016	2,458	2,879	14,720	2,039
2019-2020 Language Arts	183	1,053	1,616	1,907	2,356	11,190	2,270
2020-2021 Language Arts	891	4,024	5,584	6,396	7,654	34,140	2,394
2018-2019 Reading	153	1,635	2,858	4,025	5,132	24,567	1,937
2019-2020 Reading	144	991	1,871	2,814	3,771	18,605	2,055
2020-2021 Reading	392	3,295	5,694	8,330	10,709	59,837	2,395

Race. Because the student-level demographic data file provides very limited demographic information, the percentage values of this column were assigned to each student based on their school district (AccountID). The account-level data provides the percentages of white students in the school district. Here, the students were considered a high majority (coded as 1) district if 50% or more students in the school were white, and they were considered a low majority (coded as 0) district otherwise. Table 5 shows that nearly ⅔ of the students were from majority districts while approximately ⅓ were from non-majority districts.

TABLE 5. PERCENTAGE OF STUDENTS BY RACE, YEAR, AND SUBJECT AREA (AFTER DATA CLEANING)

Dataset	N Count	Non-majority	Majority
2018-2019 Math	201,118	36%	64%
2019-2020 Math	182,256	39%	61%
2020-2021 Math	468,112	36%	64%
2018-2019 Language Arts	120,619	34%	66%
2019-2020 Language Arts	118,406	39%	61%
2020-2021 Language Arts	344,654	35%	65%
2018-2019 Reading	172,660	32%	68%
2019-2020 Reading	157,893	39%	61%
2020-2021 Reading	455,209	35%	65%

Table 6 shows the number of items with fewer than 100 or 200 students in the focal group (non-majority) and the reference group (majority). As shown in Table 6, the 2018-2019 math item pool had 292 items where the focal group had fewer than 100 students.

TABLE 6. THE NUMBER OF ITEMS WITH FEWER THAN 100 AND 200 IN FOCAL AND REFERENCE GROUPS FOR RACE (AFTER DATA CLEANING)

Dataset	Number of Items	Focal < 100	Focal < 200	Reference< 100	Reference< 200
2018-2019 Math	3,775	292	758	3	106
2019-2020 Math	3,121	24	278	0	24
2020-2021 Math	3,380	1	22	0	6
2018-2019 Language Arts	1,970	128	154	6	116
2019-2020 Language Arts	2,302	26	118	0	60
2020-2021 Language Arts	2,395	0	0	0	0
2018-2019 Reading	2,199	1	57	4	65
2019-2020 Reading	2,383	37	116	5	68
2020-2021 Reading	2,394	0	0	0	0

Note. Non-majority is focal

Table 7 shows the distribution of sample sizes (focal and reference combined) for the DIF analyses conducted using the race variable. The minimum sample size for 2018-2019 mathematics analyses had 134 students in the combined reference and focal categories. The median sample sizes were highest for the 2020-21 school year and lowest for the 2019-2020 school year.

TABLE 7. DESCRIPTIVE SAMPLE SIZES IN DIF ANALYSIS USING RACE (AFTER DATA CLEANING)

Dataset	Minimum	1st Quarter	Median	Mean	3rd Quarter	Max	# Total Items
2018-2019 Math	134	867	3,642	4,951	7,567	29,873	3,168
2019-2020 Math	187	1,438	3,476	4,245	5,697	27,849	3,059
2020-2021 Math	243	3,941	8,056	9,904	13,172	64,348	3,380
2018-2019 Language Arts	120	2,358	3,552	4,225	5,049	25,158	2,094
2019-2020 Language Arts	167	1,894	2,982	3,455	4,243	21,283	2,289
2020-2021 Language Arts	1,284	5,838	8,068	9,277	11,153	49,223	2,394
2018-2019 Reading	166	2,589	4,720	6,568	8,732	38,496	1,941
2019-2020 Reading	152	1,703	3,356	4,870	6,686	33,128	2,116
2020-2021 Reading	557	4,774	8,258	12,088	15,498	88,027	2,395

Socioeconomic Status: The account-level data provided the percentages of children in the district from families below the poverty line. The poverty data was sourced from the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program. The poverty percentage used in this study identified districts and public schools by the actual percentage of children in the district that come from families below the poverty line. This percentage was calculated by creating a ratio of the children in a district from families below the poverty line to all children in the district. Again, the percentage values of this column were assigned to each student based on their school district. Here, students were considered a part of a high poverty district (coded as 1) if more than 17% of students were living in poverty, and they were in a low poverty district (coded as 0) otherwise. Originally, the intention was to assign high SES districts using a 50% cut off; however, there were very few districts available where more than 50% of students lived in poverty; therefore, the average percentage of students in poverty to divide the data was used. Table 8 shows that nearly 60% of students were from schools where more than 17% were classified as high-poverty schools while nearly 40% were from low poverty schools.

TABLE 8. PERCENTAGE OF STUDENTS BY SES, YEAR, AND SUBJECT AREA (AFTER DATA CLEANING)

Dataset	N Count	Low SES	High SES
2018-2019 Math	197,909	37%	64%
2019-2020 Math	169,988	44%	56%
2020-2021 Math	492,220	45%	55%
2018-2019 Language Arts	119,389	37%	63%
2019-2020 Language Arts	107,846	38%	62%
2020-2021 Language Arts	337,350	41%	59%
2018-2019 Reading	169,482	40%	60%
2019-2020 Reading	146,407	39%	61%
2020-2021 Reading	448,837	45%	55%

Table 9 shows the number of items with fewer than 100 or 200 students in the focal group (high poverty) and the reference group (low poverty). As shown in Table 9, the 2018-2019 language arts item pool had 116 items where the reference group had less than 100 students.

TABLE 9. THE NUMBER OF ITEMS WITH FEWER THAN 100 AND 200 IN FOCAL AND REFERENCE GROUPS FOR SES (AFTER DATA CLEANING)

		SES			
Subject	Number of Items	Focal < 100	Focal < 200	Reference < 100	Reference < 200
2018-2019 Math	3,775	14	264	17	198
2019-2020 Math	3,121	2	119	11	60
2020-2021 Math	3,380	1	13	3	15
2018-2019 Language Arts	1,970	2	110	116	128
2019-2020 Language Arts	2,302	0	51	53	141
2020-2021 Language Arts	2,395	0	0	0	0
2018-2019 Reading	2,199	0	17	85	136
2019-2020 Reading	2,383	6	69	54	148
2020-2021 Reading	2,394	0	0	0	1

Note. High poverty is focal.

Table 10 shows the distribution of sample sizes (focal and reference combined) for the DIF analyses conducted using the SES variable. The minimum sample size for 2018-2019 mathematics analyses had 115 students in the combined reference and focal categories. The median sample sizes were highest for the 2020-21 school year and lowest for the 2019-2020 school year.

TABLE 10. DESCRIPTIVE SAMPLE SIZES IN DIF ANALYSIS USING SES (AFTER DATA CLEANING)

Subject	Min	1st Qu	Median	Mean	3rd Qu	Max	# Total Items
2018-2019 Math	115	962	3,797	5,000	7,563	29,239	3,076
2019-2020 Math	196	1,326	3,249	3,965	5,329	26,053	3,047
2020-2021 Math	236	3,845	7,842	9,649	12,813	62,836	3,380
2018-2019 Language Arts	122	2,356	3,548	4,208	5,036	24,836	2,085
2019-2020 Language Arts	158	1,746	2,706	3,150	3,833	19,514	2,273
2020-2021 Language Arts	1,262	5,720	7,909	9,075	10,851	48,773	2,394
2018-2019 Reading	151	2,483	4,573	6,410	8,499	38,016	1,948
2019-2020 Reading	100	1,599	3,141	4,535	6,235	30,579	2,113
2020-2021 Reading	544	4,701	8,133	11,926	15,388	86,651	2,395

Pandemic. The pandemic grouping variable was obtained by appending the pre-pandemic data (all items administered prior to March 2020) to the pandemic data (all items administered after March 2020). The pre-pandemic data combined data from the 2018-2019 and 2019-2020 data sets while the pandemic data combined any responses from 2019-2020 administered after March 2020 with the 2020-2021 data. Table 11 shows that nearly 40% of the cases were from the pre-pandemic era while nearly 60% were from the pandemic era. There were no items with fewer than 200 responses in pre- and post-pandemic groups.

TABLE 11. PERCENTAGE OF STUDENTS BY PANDEMIC AND SUBJECT AREA (AFTER DATA CLEANING)

Subject	N Count	Pre-Pandemic	Pandemic
Math	966,420	42%	58%
Language Arts	677,223	38%	62%
Reading	897,623	39%	61%

Table 12 shows the distribution of sample sizes (focal and reference combined) for the DIF analyses conducted using the pandemic variable. Here, the minimum sample was over 1,000 students for each content area.

TABLE 12. DESCRIPTIVE STATISTICS OF THE RESPONSES IN DIF ANALYSIS USING PANDEMIC (AFTER DATA CLEANING)

Subject	Min	1st Qu	Median	Mean	3rd Qu	Max	# Total Items	# Items with N < 400	% Items with N < 400
Math	1,187	8,469	18,709	21,779	29,416	117,827	2,778	0	0.00%
Language Arts	2,598	12,601	17,096	19,228	22,799	96,758	2,090	0	0.00%
Reading	1,861	12,339	20,886	28,221	36,378	162,735	1,865	0	0.00%

RESULTS

The results for each grouping variable are presented in this section. Only A-level DIF was identified in these analyses for every grouping variable. The findings did not indicate B- or C-level DIF.

GENDER-RELATED DIF

Table 13 shows the results for gender-related DIF including the total number of unique items after applying the cleaning rules, the number of unique items flagged for A-level DIF, the number of DIF items flagged in favor of the focal group (males), and the number of items flagged in favor of the reference group (females). Again, no items were flagged for B- or C-level gender-related DIF. Table 13 shows that very few items were flagged for DIF across all years and subject areas. The largest number of DIF items

was found in the 2019 reading data where nearly 7% of the items were flagged for A-level DIF. Further investigation showed that most of the flagged items were flagged in favor of females.

TABLE 13. NUMBER OF ITEMS CLASSIFIED WITH A-LEVEL GENDER-RELATED DIF BY YEAR AND SUBJECT AREA (AFTER DATA CLEANING)

Dataset	Number of Unique Items	Number of Items Flagged for A-Level DIF	Number of Items Flagged in Favor of Focal Group	Number of Items Flagged in Favor of Reference Group	Number of Items Not Flagged
2018-2019 Math	2,932	3	0	3	2,929
2019-2020 Math	3,040	3	0	3	3,037
2020-2021 Math	3,377	14	3	11	3,363
2018-2019 Language Arts	2,039	98	11	87	1,941
2019-2020 Language Arts	2,270	56	8	48	2,214
2020-2021 Language Arts	2,394	114	23	91	2,280
2018-2019 Reading	1,937	134	51	83	1,803
2019-2020 Reading	2,055	74	31	43	1,981
2020-2021 Reading	2,395	142	59	82	2,254

RACE-RELATED DIF

Table 14 shows the results for race-related DIF including the total number of unique items after applying the cleaning rules, the number of unique items flagged for A-level DIF, the number of DIF items flagged in favor of the focal group (non-majority), and the number of items flagged in favor of the reference group (majority). Again, no items were flagged for B- or C-level race-related DIF. Table 14 shows more items were flagged for A-level race-related DIF than for gender-related DIF. Over 10% of the items were flagged in language arts in 2019, reading in 2019, and reading in 2021. In most cases, most of the items were flagged in favor of the majority group.

TABLE 14. NUMBER OF ITEMS CLASSIFIED WITH A-LEVEL RACE-RELATED DIF BY YEAR AND SUBJECT AREA (AFTER DATA CLEANING)

Dataset	Number of Unique Items	Number of Items Flagged for A-Level DIF	Number of Items Flagged in Favor of Focal Group	Number of Items Flagged in Favor of Reference Group	Number of Items Not Flagged
2018-2019 Math	3,168	144	34	110	3,024
2019-2020 Math	3,059	149	58	91	2,910
2020-2021 Math	3,380	74	10	64	3,306
2018-2019 Language Arts	2,094	96	32	64	1,998
2019-2020 Language Arts	2,289	193	111	82	2,096
2020-2021 Language Arts	2,394	168	60	108	2,226
2018-2019 Reading	1,941	177	68	109	1,764
2019-2020 Reading	2,116	246	110	136	1,870
2020-2021 Reading	2,395	218	77	141	2,177

SES-RELATED DIF

Table 15 shows the results for SES-related DIF including the total number of unique items after applying the cleaning rules, the number of unique items flagged for A-level DIF, the number of DIF items flagged in favor of the focal group (high poverty), and the number of items flagged in favor of the reference group (low poverty). Again, no items were flagged for B- or C-level SES-related DIF. Table 15 shows fewer items were flagged for A-level SES-related DIF than for race-related DIF. Here, over 10% of the items were flagged in reading in 2020. In most cases, most items were flagged in favor of the low-poverty group.

TABLE 15. NUMBER OF ITEMS CLASSIFIED WITH A-LEVEL SES-RELATED DIF BY YEAR AND SUBJECT AREA (AFTER DATA CLEANING)

Dataset	Number of Unique Items	Number of Items Flagged for A-Level DIF	Number of Items Flagged in Favor of Focal Group	Number of Items Flagged in Favor of Reference Group	Number of Items Not Flagged
2018-2019 Math	3,076	188	140	48	2888
2019-2020 Math	3,047	168	105	63	2879
2020-2021 Math	3,380	3	3	0	3377
2018-2019 Language Arts	2,085	232	149	83	1853
2019-2020 Language Arts	2,273	203	86	117	2070
2020-2021 Language Arts	2,394	181	121	60	2213
2018-2019 Reading	1,948	294	208	86	1654
2019-2020 Reading	2,113	203	104	99	1910
2020-2021 Reading	2,395	376	260	116	2019

PANDEMIC-RELATED DIF

Table 16 shows the results for the pandemic-related DIF including the total number of unique items after applying the cleaning rules, the number of unique items flagged for A-level DIF, the number of DIF items flagged in favor of the focal group (post-pandemic), and the number of items flagged in favor of the reference group (pre-pandemic). Again, no items were flagged for B- or C-level pandemic-related DIF. Table 16 shows nearly ⅓ of the items were flagged for A-level DIF, and in most cases, most items were flagged in favor of the post-pandemic period.

TABLE 16. NUMBER OF ITEMS CLASSIFIED WITH A-LEVEL PANDEMIC-RELATED DIF BY YEAR AND SUBJECT AREA (AFTER DATA CLEANING)

Dataset	Number of Unique Items	Number of Items Flagged for A-Level DIF	Number of Items Flagged in Favor of Focal Group	Number of Items Flagged in Favor of Reference Group	Number of Items Not Flagged
Math	2,778	943	561	382	1835
Language Arts	2,090	733	564	169	1357
Reading	1,865	716	509	207	1149

CONCLUSIONS

The primary purpose of this study was to investigate the Edmentum's item bank for gender, SES, race, and pandemic-related DIF. When conducting DIF studies with the ETS classification system, items were classified as A-, B-, or C-level DIF. Items classified with A-level DIF have "little or no difference between the two matched groups" (Zieky, 2003). Items flagged with B- and C-level DIF were the items of concern. Despite the large number of items in the Edmentum's item bank, no items were flagged for B- or C-level DIF.

This finding was surprising given the large number of available items. The MH procedure is a robust procedure for detecting uniform DIF but not non-uniform DIF. It may be that non-uniform DIF has gone undetected in the analyses. Additionally, the student-level data does not contain the demographic information beyond gender; thus, school-level data was applied to investigate race- and SES-related DIF. This type of categorization may have masked differences in performance that may have been uncovered with more refined data.

DIF procedures are one means of examining test construct. When tests are beset with items flagged for DIF, this indicates that the construct may be measured differently between groups. The lack of DIF items in the current study provides evidence that the items in Edmentum's item bank are measuring student achievement from different groups in a similar manner. It provides some evidence that Edmentum's items are fair for different groups.

REFERENCES

Clauser, B.E., & Mazor, K.M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*, 31-44.

Cohen, J. (1988). *Statistical Power Analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A programmatic perspective. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135-165). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fidalgo, Á. M., Ferreres, D., & Muñiz, J. (2004). Liberal and conservative differential item functioning detection using Mantel-Haenszel and SIBTEST: Implications for Type I and Type II error rates. *The Journal of Experimental Education*, *73*(1), 23-39.

Fidalgo, Á. M., Mellenberg, G. J., & Muñiz, J. (2000). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure and the iterative logit method. *Revista Electrónica de Metodología Aplicada, 5*(1), 10-20.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*(4), 278-295.

Gierl,M. J., Jodoin, M. G., & Ackerman, T. A. (2000). *Performance of Mantel-Haenszel, simultaneous item bias test, and logistic regression when the proportion of DIF items is large*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana.

Güler, N., & Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. J*ournal of Educational Measurement*, *46*(3), 314-329.

Holland, P. W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.) *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Huang, C. D., Church, A. T., and Katigbak, M. S. (1997). Identifying culturual differences in items and traits: Differential item functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology*, *28*, 192-218.

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*, 847-862.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel_Haenszel Procedure. *Educational and Psychological Measurement*, *54*, 284-291.

Rogers, H. J., & Swaminathan, H. (1993). A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning. *Applied Psychological Measurement*, *17*(2), 105-116.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement*, *33*(2), 215-230.

Wiberg, M. (2009). Differential item functioning in mastery tests: A comparison of three methods using real data. *International Journal of Testing*, *9*, 41-59.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4, 1686.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics*, *15*(3), 185–197.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, *26*, 55-66.